Udacity Data Engineering Nanodegree Capstone Project

John Maxi

Step 1: Scope the Project and Gather Data

The goal of my project is to build an analytics table for researching the relationship between air quality and person-level health outcomes. Air quality data will be obtained from the EPA <https://aqs.epa.gov/aqsweb/airdata/download_files.html> and person-level health outcome will be obtained from the Nation Health Interview Survey <https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm>.

Step 2: Explore and Assess the Data

Step 3: Define the Data Model

Air quality data: The primary key will be the daily timestamp, the columns will indicate the different measured values for each day: PM2.5, pressure, temp, humidity, wind. There are also going to be many locations per variable, so I will need to take that into account as well. Key will need to end up including both the time and space distribution of the readings.

Should I have 2 fact tables or just one? For an “analytics table” use case I should distill all the information into a single star schema. For a “source-of-truth” database I think I want to keep the raw data, so I would have the air data table with dimensions and the health data table with dimensions. Would I be joining them with the dimensions?

A monitor is indicated by the site (state + county + site number), pollutant code and POC

PM 2.5:

State Code: Int, range 1-80

County Code: Int, range 1-810

Site Num: Int, range 1-9997

Parameter Code: Int, same value for all rows in a yearly file

POC: Int, range 1-33. POC stands for parameter occurrence code. Used to uniquely identify a monitor if there is more than one device measuring the same pollutant at the same site.

Latitude: numeric

Longitude: numeric

Datum: varchar, NAD83 or WGS84. What is this? It is associated with the latitude and longitude

Parameter: varchar, same for all rows of a yearly PM2.5 file 2.5 – Local Conditions. I think this probably corresponds to the Parameter column. Probably each parameter matches the data type per file (so humidity has a different parameter number).

Sample Duration: varchar, 1 HOUR, 24 HOUR, 24-HR BLK AVG. This refers to the length of time air flows though a monitor before being analyzed. It represents an averaging period.

Pollutant Standard: varchar, PM25 24-hour 2012 or blank. This indicates how the pollutant values are calculated. The EPA standard is 24 hour averages. Some of the rows are 1 hour duration averages

Date Local: date

Units of Measure: varchar

Event type: Indicates if data is included after an uncontrollable event occurred (e.g., wildfire)

Observation Count: int, number of observations taken that day

Observation Percent: int, number of observations taken with respect to the number of scheduled for that day

Arithmetic mean: numeric, the average value for the day

1st Max Value: numeric, the highest value for the day. Only meaningful when the sample duration is 1 hour

1st Max Hour: int, the hour at which the 1st Max Value occurred. Is always 0 in the daily summery. Only meaningful when the sample duration is 1 hour

AQI: Air Quality Index for the day

Method code: int, internal code for identifying the processes/protocol of data collection

Method name: varchar

Local Site Name: varchar, name of the site given by local operators

Address: varchar

State Name: varchar

County: varchar

City Name: varchar

CBSA Name: varchar, core bases statistical area (metropolitan area where the monitoring site is located

Date of Last Change: date, the last time any numeric value was changed in the record

I’m thinking of a data model that splits this table into a fact table – readings, and dimension tables for time and place aspects of the data. At this point I’m planning to only keep the 24 data so I won’t need the 1st max or 1st max hour columns at all.